第 43 卷 第 12 期/2023 年 6 月/光学学报 研究论文

光学学报

支持向量机在混合气体定量分析中的应用

闪霁芳[1,2]，刘琨[1,2*]，江俊峰[1,2]，刘铁根[1,2]，尹慧[1,2]

[1]天津大学精密仪器与光电子工程学院，天津 300072；

[2]天津大学光电信息技术教育部重点实验室，天津 300072

摘要 针对使用掺铥光纤激光器的气体传感系统进行混合气体测量时，吸收谱线重叠较为严重且相互交叉吸收干扰的现象造成的测量误差大、分析精度低的问题，提出一种基于自适应变异粒子群优化的支持向量机(SVM)方法，用于建立混合气体体积分数定量分析预测模型。对体积分数为0.5%～2%的氨气(NH3)和2%～5%的二氧化碳(CO_2)混合气体的吸收光谱数据进行采集和处理，利用自适应变异粒子群优化(AMPSO)算法对SVM模型参数进行寻优，利用获得的最优模型参数构建氨气和二氧化碳气体体积分数定量分析模型，并与标准粒子群优化算法和网格搜索法进行对比。实验结果表明，基于自适应粒子群优化算法建立的氨气和二氧化碳气体体积分数定量分析模型在较为合适的寻优时间下，可以得到最佳的均方误差，效率较高，该模型对测试集中氨气和二氧化碳气体体积分数设定值与预测值的均方误差分别为0.000088和0.000170，决定系数R^2均为0.9998，满足混合气体检测要求。

关键词 光通信；掺铥光纤激光器；自适应变异粒子群优化；混合气体；支持向量机

中图分类号 TN29 **文献标志码** A **DOI**：10.3788/AOS221681

1 引 言

机动车尾气含有氨气(NH_3)、二氧化碳(CO_2)等气体，成为大气污染和温室效应的重要来源[1]。基于光纤环腔激光器的内腔吸收气体传感技术具有响应速度快、光谱覆盖范围大、光谱分辨率高与检测灵敏度高等优点，且通过构建掺铥光纤激光器可将增益带宽覆盖至具有多种污染气体强吸收的2 μm波段，该技术非常适合用于环境保护领域对有毒有害气体的实时检测[2-5]。

基于掺铥光纤激光器的气体传感系统进行混合气体的定量分析时，组分气体间吸收谱线的重叠产生交叉干扰以及实验现场温度压强改变导致的非线性偏移等因素[6]常常会影响气体检测精度。近年来，利用化学计量方法与红外光谱分析相结合建立的混合气体浓度回归预测模型，对非线性干扰进行修正，可极大地提高气体定量分析的准确性和可靠性，常见的修正算法有人工神经网络(ANN)[7-9]、支持向量机(SVM)[10-12]等。Zhang等[13]使用线性判别分析结合ANN算法对人工鼻传感器采集的NH_3气体进行浓度响应，对于不同浓度的NH_3样本的最大误差为5.60%；Soroush等[14]使用列文伯格-马夸尔特法、反向传播算法结合ANN算法对溶液中CO_2浓度建立回归分析网络，预测浓度的决定系数R^2为0.9828。然而，神经网络非线性模型可能会受到局部极小、过拟合、复杂结构和随机初始权值等因素的影响[11,15]。相较于上述两种方法，基于统计学理论的SVM作为一种小样本机器学习方法，通过引入核函数，有效避免了维数灾难和局部极小问题，该方法具有较高的准确率和良好的泛化能力，是一种较为准确的混合气体浓度回归预测方法。

本文通过自行搭建的基于掺铥光纤激光器的气体传感系统，对体积分数为0.5%～2%的NH_3和体积分数为2%～5%的CO_2混合气体的吸收光谱数据进行采集，通过预处理和主成分分析法进行光谱信息特征提取后，使用自适应变异粒子群优化算法对SVM模型进行参数优化，通过训练得到NH_3和CO_2气体体积分数回归预测模型，分别为NH_3-SVM model和CO_2-SVM model。实验结果证明了使用SVM算法模型进行自适应变异粒子群优化在混合气体定量分析应用中的可行性。

2 基本原理

2.1 内腔吸收气体传感技术原理

待测气体对红外光的吸收依据是朗伯-比尔定

收稿日期：2022-09-06；**修回日期**：2022-10-13；**录用日期**：2022-10-27；**网络首发日期**：2022-11-04

基金项目：国家自然科学基金(61922061,61735011,61775161)、国家重大仪器设备开发专项(2013YQ 030915)、天津市自然科学基金杰出青年科学基金(19JCJQJC61400)

通信作者：*beiyangkl@tju.edu.cn

律[16]。当一束红外光通过气体介质时，入射光强度$I_0(v)$和出射光强度$I(v)$的关系可表示为

$$I(v)=I_0(v)\exp\left[-\alpha(v)cL\right], \tag{1}$$

式中：v为激光频率；c为气体体积分数；$\alpha(v)$为气体的吸收系数（其大小与c无关，只与气体种类和特定频率有关）；L为待测气体与入射光相互作用的距离，也称有效吸收光程。

为了表征不同种类、不同体积分数气体吸收的强弱，定义吸光度$K(v)$表示待测气体吸收导致的系统光强衰减，对式(1)变形可得吸光度$K(v)$的表达式为

$$K(v)=\ln\frac{I_0(v)}{I(v)}=\alpha(v)cL。 \tag{2}$$

由式(2)可知，当温度、压强和气体有效吸收光程一定时，理想状态下气体体积分数和吸光度呈正比关系，通过对待测气体体积分数和吸光度进行体积分数标定，就可以实现待测气体体积分数的定量分析。

在搭建的掺铥光纤激光器中，将装有待测气体的传感气室放入激光器谐振腔内，构成掺铥光纤环腔传感系统，如图1所示。这种直接吸收式的传感方式无需加入调制信号，传感结构简单，操作方便。但是实际测量会因为以下因素产生偏差：1）混合气体在邻近波长位置的吸收峰重叠，产生交叉干扰现象；2）吸光度很容易受到实验现场的温度和压强的影响，导致正比关系偏移；3）系统的入射光源不是理想状态下的单色光，其光谱有一定的宽度，待测气体对不同频率红外光的吸收能力不同，导致吸收谱线产生一定的展宽，影响气体吸收峰的定位[6]。

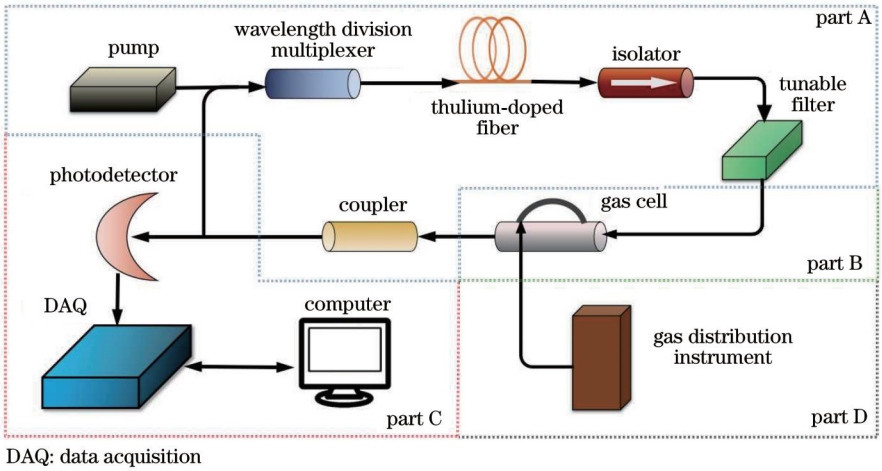

图1　基于掺铥光纤环腔传感采集系统的原理图

Fig. 1　Schematic of sensing acquisition system based on thulium-doped fiber ring cavity

2.2　SVM回归预测模型的基本原理

SVM模型结构是由Vapnik在20世纪60年代提出的[17]，它是一种建立在统计学理论和结构风险最小化原则上的新一代机器学习系统，具有泛化能力强和结构简单等优点[18]，被广泛应用于不同领域的预测回归。

假设气体光谱训练数据集为$\left\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n)\right\}$，其中$i=1,2,\cdots,n$，$n$为数据集中的样本总数，$\boldsymbol{x}_i\in\mathbf{R}^p$为一个$p$维的输入向量，代表第$i$个样本的光谱数据，$y_i$为与之对应的第$i$个样本的气体体积分数。设拟合回归函数为

$$f(x)=\boldsymbol{w}\boldsymbol{x}+b, \tag{3}$$

式中：$\boldsymbol{w}$为一个权矢量，垂直于最优回归超平面；b为偏差项；$\boldsymbol{x}$为气体光谱样本数据。

引入ε-不敏感损失函数，假设所有样本点$\boldsymbol{x}_i$在精度ε下均可以无误差地拟合成$y_i=f(\boldsymbol{x}_i)$，则有约束条件：$\left|y_i-f(\boldsymbol{x}_i)\right|\leqslant\varepsilon$。

为了保证回归函数一定有解，引入弛豫变量ξ、ξ^*，分别对应数据点允许向上和向下偏离目标函数的量，则优化问题变为求解最小值问题，即

$$\min_{w,\xi,\xi^*}\frac{1}{2}\|\boldsymbol{w}\|^2+C\left(\sum_{i=1}^{n}\xi_i+\sum_{i=1}^{n}\xi_i^*\right)$$
$$\text{s.t.}\begin{cases}y_i-(\boldsymbol{w}\boldsymbol{x}_i+b)\leqslant\varepsilon+\xi_i\\(\boldsymbol{w}\boldsymbol{x}_i+b)-y_i\leqslant\varepsilon+\xi_i^*\\\xi_i,\xi_i^*\geqslant0\end{cases}, \tag{4}$$

式中：C为惩罚因子，代表模型对误差的容忍度，C越大表示对训练误差大于损失函数ε的样本惩罚越大。

通过引入拉格朗日乘子α_i、α_i^*、β_i、β_i^*（α_i、α_i^*、β_i、β_i^*都不小于0），可以将上述有约束的原始函数转化成无约束的新构造的拉格朗日函数，其对偶形式如下：

$$L=\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{n}(\xi_i+\xi_i^*)-\sum_{i=1}^{n}\alpha_i[\xi_i+\varepsilon-y_i+(\boldsymbol{w}\boldsymbol{x}_i+b)]-$$
$$\sum_{i=1}^{n}\alpha_i^*[\xi_i^*+\varepsilon+y_i-(\boldsymbol{w}\boldsymbol{x}_i+b)]-\sum_{i=1}^{n}(\beta_i\xi_i+\beta_i^*\xi_i^*)。\tag{5}$$

根据边界条件，拉格朗日最优化的对偶问题为

$$\max_{\alpha_i,\alpha_i^*}\sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*)-\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)-\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(\boldsymbol{x}_i,\boldsymbol{x}_j)\ \text{s.t.}\begin{cases}\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)=0\\0\leqslant\alpha_i,\alpha_i^*\leqslant C\end{cases}。\tag{6}$$

式(6)中拉格朗日乘子α_i、α_i^*不全为0时对应的样本点$\boldsymbol{x}_i$称为支持向量。任取一组支持向量组合$(\boldsymbol{x}_i,y_i)$，在约束边界条件下，根据优化解的性质(Karush-Kuhn-Tucker conditions)可求得偏置项b，最后计算得到的预测输出为

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\boldsymbol{x}_i\boldsymbol{x}+b。\tag{7}$$

对于混合气体而言，原始样本空间是一个非线性体系，通过引入核函数$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$，就可以实现从低维的样本空间非线性到高维特征空间的变换并进行线性求解。当核函数$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$符合Mercer定理时，它就可以对应一个变换空间$\mathbf{R}^p$中的一个内积，则混合气体条件下非线性问题中的回归函数为

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\kappa(\boldsymbol{x}_i,\boldsymbol{x})+b。\tag{8}$$

选用不同的核函数可以构造出不同类型的SVM模型，常见的核函数有线性核函数、多项式核函数、径向基核函数和Logistic核函数。通过实验和经验分析可得，对于混合气体模型，本文采用以下高斯径向基函数(RBF)来构造SVM模型，即

$$\kappa(x,y)=\exp\left(-g\|x-y\|^2\right),\tag{9}$$

式中：g为核函数参数，$g>0$。

3 实验系统

本文自行搭建了基于掺铥光纤激光器的有源内腔气体传感系统，用于采集NH_3和CO_2气体的吸收光谱数据。系统原理图如图1所示，主要分为可调光源部分(part A)、传感部分(part B)、数据采集处理部分(part C)和配气部分(part D)，实物图如图2所示。

可调光源部分包括1570 nm波长的泵浦光源、1550 nm/2000 nm波分复用器、隔离器、可调谐滤波器(调谐范围为1928.5～2053.0 nm)、5 m单模掺铥光纤(TDF)和10/90耦合器(90%的输出端连接波分复用器，将绝大部分的光返回环腔中)。以上器件按照一定顺序相互连接，形成环形系统，通过可调谐滤波器的选频作用，在谐振腔内形成激光振荡，获得1928.5～1985.0 nm范围可调谐的宽带放大自发辐射光输出。当泵浦光功率为1.2 W时，系统输出光谱如图3所示，

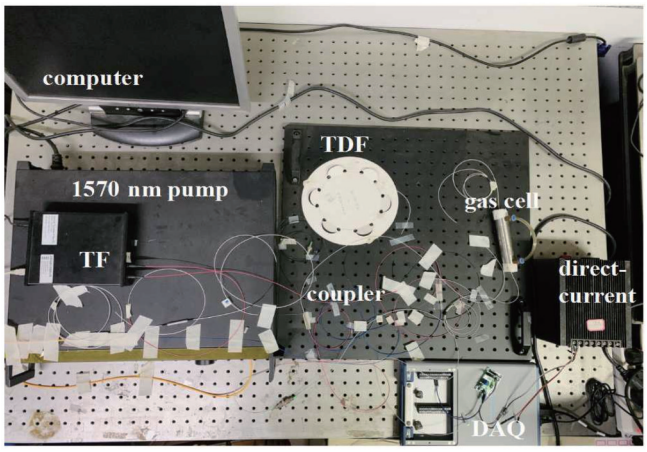

图2 基于掺铥光纤环腔的传感采集系统实物图

Fig. 2 Physical map of sensing acquisition system based on thulium-doped fiber ring cavity

可以看到，所搭建的环腔系统可以实现56 nm的平坦输出。

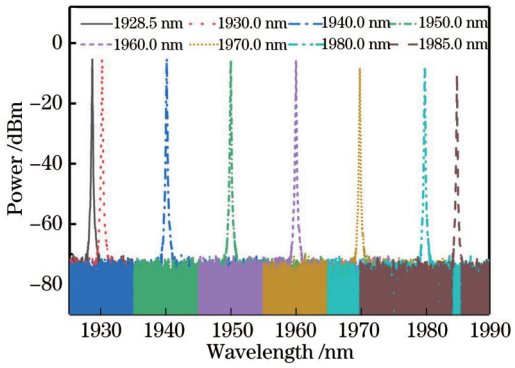

图3 基于掺铥光纤环腔的系统的输出光谱图

Fig. 3 Output spectra of thulium-doped fiber ring cavity system

传感部分为自制的传感气室，将气室设计成上端带有进气口和出气口、两端装有C-lens准直器的空心钢管结构，方便待测气体的通入和废弃气体的排出。输出的自发辐射光通过装有待测气体的传感气室，引起待测气体对特定波长光的吸收，产生光强的衰减，环腔结构使得光信号在环路可以多次通过气室，大大增加了有效吸收光程；数据采集处理部分包括InGaAs/PIN近红外光电探测器、NI数据采集卡和计算机，实现对带有体积分数信息的光信号的转换和电

信号的采集处理；配气部分为SY-9506型配气仪，在常压下2%标准体积分数的NH_3和5%标准体积分数的CO_2气体分别通过质量流量计混合后通入气室，通过设置NH_3和CO_2气体的质量配比，根据对应气体的质量流量计的读数就可得到混合气体中NH_3和CO_2气体的体积分数。

在采集气体光谱前，向气室内通入足量的氮气以排除气室中水蒸气和CO_2的干扰。本文实验环境为0.1 MPa常压，采集卡的采样率为20 kHz，共采集了20组数据，每组数据包含12个样本，如表1所示，其中编号1～8是体积分数为0.5%～2%的96个纯NH_3气体样本数据，编号9～12是体积分数为2%～5%的48个纯CO_2气体样本数据，编号13～20是96个NH_3和CO_2混合气体样本数据。

表1 训练集数据

Table 1 Data of train set

Group	NH_3 volume fraction /%	CO_2 volume fraction /%	Group	NH_3 volume fraction /%	CO_2 volume fraction /%
1	0.5	0	11	0	4.0
2	1.0	0	12	0	5.0
3	1.2	0	13	1.0	2.0
4	1.4	0	14	2.0	2.0
6	1.6	0	16	2.0	3.0
7	1.8	0	17	1.0	4.0
8	2.0	0	18	2.0	4.0
9	0	2.0	19	1.0	5.0
10	0	3.0	20	2.0	5.0

4 实验结果与分析

4.1 光谱数据预处理

在搭建模型前，对光谱数据进行一定的预处理，以减小背景噪声的影响，提高信噪比，但是不宜进行过多的预处理，避免丢失某些重要的光谱信息。本文对采集的光谱数据进行了去噪、基线校正和平滑处理。

在采集光谱信号时，采用LabVIEW软件编写低通滤波程序，对高频的噪声信号进行过滤，得到去噪后的吸收光谱图。将光谱数据导入Origin软件，采用人工基线校正的方法得到基线在0位置的光谱图，使用Savitzky-Golay卷积平滑算法对最终光谱进行平滑处理，消除噪声和漂移等因素对光谱数据的影响，提高体积分数分析的准确率。由于实验整体在常压下进行，气体谱线展宽主要为碰撞展宽，因此采用洛伦兹线型对吸收谱线进行拟合。图4(a)～(c)分别为体积分数为2%的纯NH_3气体经预处理后的吸收光谱数据图、经过洛伦兹拟合后的吸收光谱数据图以及根据HITRAN数据库仿真得到的吸收光谱数据图；图4(d)～(f)分别为体积分数为5%的纯CO_2气体经预处理后的吸收光谱数据图、经过洛伦兹拟合后的吸收光谱数据图以及根据HITRAN数据库仿真得到的吸收光谱数据图；图4(g)～(i)分别为体积分数为2%的NH_3和体积分数为2%的CO_2混合气体经预处理后的吸收光谱数据图、经过洛伦兹拟合后的吸收光谱数据图以及根据HITRAN数据库仿真得到的吸收光谱数据图。可以看到，实验采集处理后的气体吸收光谱数据经过洛伦兹拟合后，气体吸收峰与HITRAN理论数据吻合得较好，这验证了使用本系统采集气体吸收光谱数据的可行性，同时也可看出混合气体吸收峰存在明显的重叠现象，需要建立体积分数回归模型进行气体体积分数回归。

4.2 主成分分析法降维

为提高建模速度，利用主成分分析(PCA)法[19]将原始气体吸收光谱数据的多维特征值进行线性变换并投影到高维空间，获取最大方差对应的主成分，并利用此时的主成分代替原始数据中的特征值，以降低数据维数，防止变量间的相关性影响主成分的提取和回归模型的预测精度。

为了得到高预测精度的回归模型，采取不同种类的气体进行独立建模。选取编号为1～8和13～20的16组数据、每组数据中的10个样本共计160个样本作为NH_3气体回归模型的训练集，剩余32个样本作为测试集，记为NH_3-SVM model；同理，选取编号为9～20的12组数据、每组数据中的10个样本共计120个样本作为CO_2气体回归模型的训练集，剩余24个样本作为测试集，记为CO_2-SVM model。此时NH_3气体回归模型的训练集矩阵大小为160×350，CO_2气体回归模型的训练集矩阵大小为120×350，对两种模型的训练集进行PCA降维处理后，得到部分主成分分布图，如图5所示。以NH_3-SVM model为例，数据降维后前15个主成分的累积贡献率高达98%，认为此时前15个主成分可以代替350个特征值所代表的信息量，数据从350维降至15维，得到得分矩阵的维度为160×15，使用降维后的得分矩阵来训练NH_3-SVM model回归模型。

4.3 AMPSO模型建立

SVM模型性能的好坏主要与RBF核函数参数g和惩罚因子C这两个参数有关，参数的选择目前没有

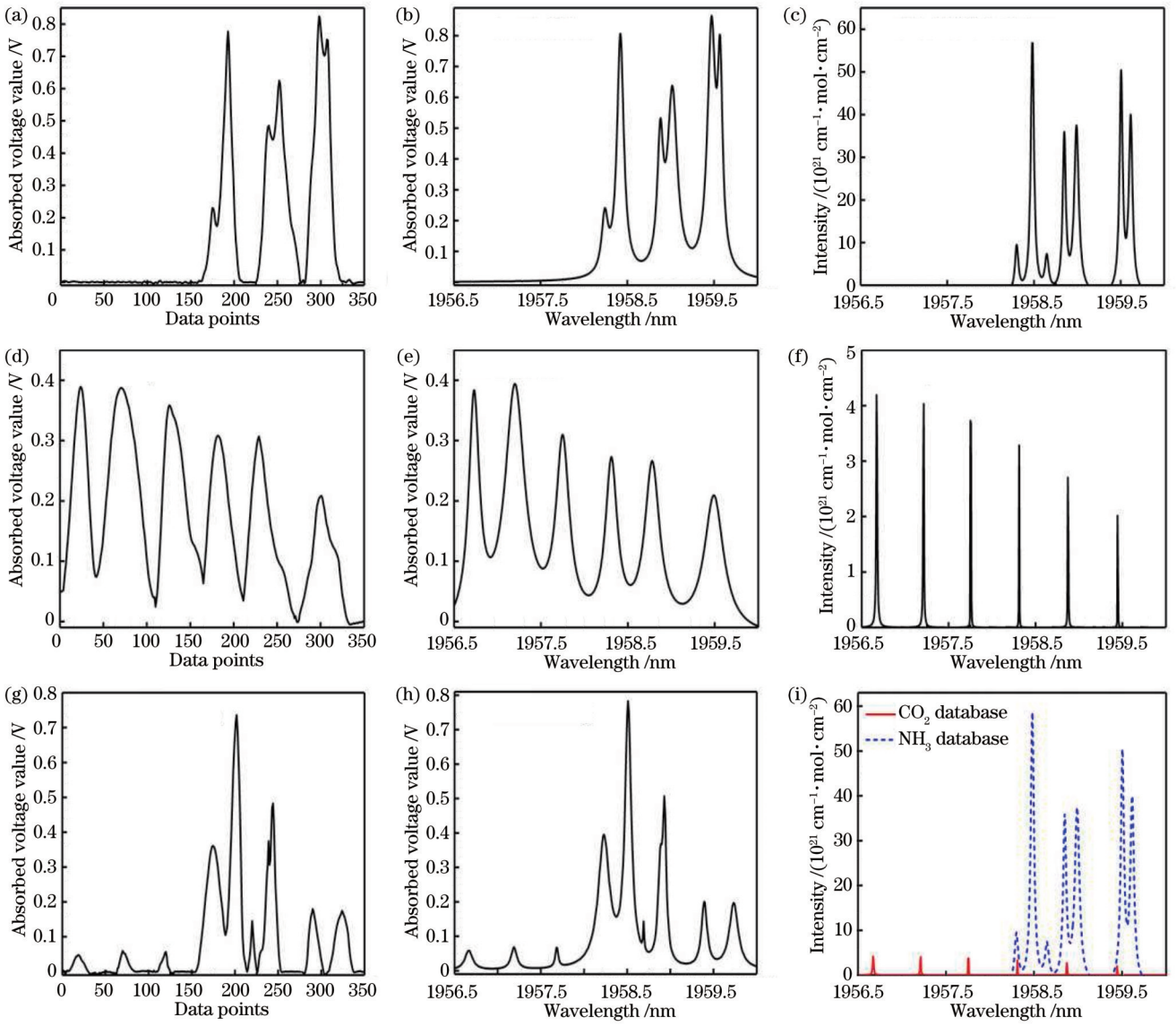

图4 吸收光谱数据示例图。(a)~(c)体积分数为2%的纯NH_3气体经预处理后的吸收光谱数据图、经洛伦兹拟合后的吸收光谱数据图以及根据HITRAN数据库仿真得到的吸收光谱数据图;(d)~(f)体积分数为5%的纯CO_2气体经预处理后的吸收光谱数据图、经洛伦兹拟合后的吸收光谱数据图以及根据HITRAN数据库仿真得到的吸收光谱数据图;(g)~(i)体积分数为2%的NH_3和体积分数为2%的CO_2混合气体经预处理后的吸收光谱数据图、经洛伦兹拟合后的吸收光谱数据图以及根据HITRAN数据库仿真得到的吸收光谱数据图

Fig. 4 Sample diagrams of absorption spectrum data. (a)-(c) Absorption spectrum data after pretreatment of pure NH_3 gas with volume fraction of 2%, absorption spectrum data after Lorentz fitting, and absorption spectrum data obtained by simulation according to HITRAN database; (d)-(f) absorption spectrum data after pretreatment of pure CO_2 gas with volume fraction of 5%, absorption spectrum data after Lorentz fitting, and absorption spectrum data obtained by simulation according to HITRAN database; (g)-(i) pre-processed absorption spectrum data of mixed gas of NH_3 with volume fraction of 2% and CO_2 with volume fraction of 2%, absorption spectrum data after Lorentz fitting, and absorption spectrum data obtained by HITRAN database simulation

标准的方法,一般要根据具体模型的多次实验决定。粒子群优化算法作为一种基于速度和位置两个随机变量搜索的智能优化方法,常用于对SVM模型参数的寻优,其物理意义是通过群体中各个个体之间的信息传递及信息共享来寻找最优解[20],在迭代过程中,粒子经个体最优解和全局最优解更新自己的速度和位置,公式为

$$\boldsymbol{V}_{id}^{(k+1)}=\omega\boldsymbol{V}_{id}^{(k)}+c_1r_1\left[\boldsymbol{P}_{id,\mathrm{pbest}}^{(k)}-\boldsymbol{X}_{id}^{(k)}\right]+c_2r_2\left[\boldsymbol{P}_{d,\mathrm{gbest}}^{(k)}-\boldsymbol{X}_{id}^{(k)}\right],\tag{10}$$

$$\boldsymbol{X}_{id}^{(k+1)}=\boldsymbol{X}_{id}^{(k)}+\boldsymbol{V}_{id}^{(k+1)},\tag{11}$$

式中:i为粒子群规模,$i=1,2,\cdots,N$;d为粒子维度,$d=1,2,\cdots,D$;k为迭代次数;ω为惯性权重;c_1为个体学习因子;c_2为群体学习因子;r_1、r_2为区间$[0,1]$内的随机数,用于增加搜索的随机性;$\boldsymbol{V}_{id}^{(k)}$为粒子$i$在第$k$次

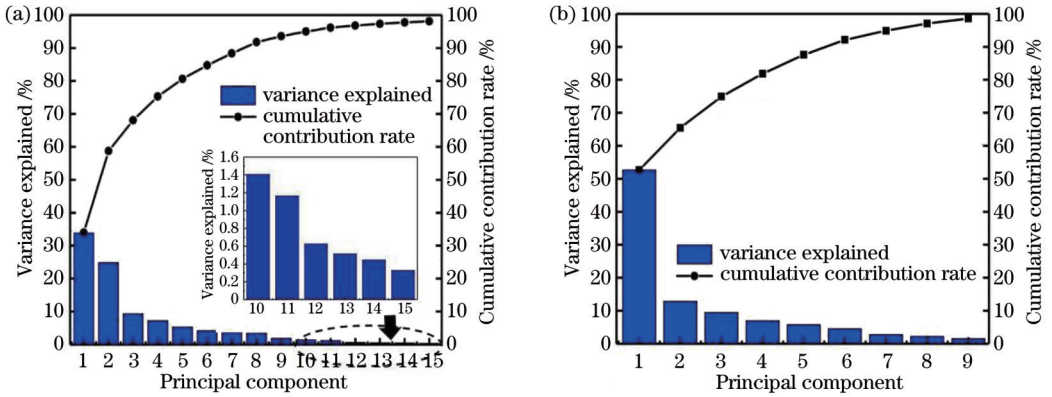

图 5　对训练集进行 PCA 降维处理后的部分主成分分布图。(a) NH_3-SVM model; (b) CO_2-SVM model

Fig. 5　Partial principal component distributions after PCA dimensionality reduction on training set. (a) NH_3-SVM model; (b) CO_2-SVM model

迭代中第 d 维的速度向量；$\boldsymbol{X}_{id}^{(k)}$ 为粒子 i 在第 k 次迭代中第 d 维的空间向量；$\boldsymbol{P}_{id,\mathrm{pbest}}^{(k)}$ 为第 i 个粒子搜索得到的最优解；$\boldsymbol{P}_{d,\mathrm{gbest}}^{(k)}$ 为整个粒子群体中的最优解。

标准 PSO 算法收敛快、优化时间短，但存在模型易早熟收敛、最优解搜索精度较低、后期迭代效率不高等问题。针对这些问题，本文提出一种改进算法——自适应变异粒子群优化（AMPSO）算法，通过引入自适应变异算子 p_{m}，对更新后的粒子位置 $\boldsymbol{X}_{id}^{(k)}$ 进行随机变异，使粒子可以进入解空间的其他区域继续进行搜索，进而提高粒子群算法跳出局部最优解的能力，避免算法模型早熟收敛。

假设粒子 i 在第 k 次迭代中的位置 $\boldsymbol{X}_{id}^{(k)}=\mathrm{pop}(j,k)$，$p_{\mathrm{m}}$ 的变异阈值为 $[0,1]$ 区间的随机常数。当粒子大于阈值时，跳出当前位置进入新的区域，否则保持不变。算法程序实现方法为

$$K=\begin{cases}\mathrm{ceil}(2r_1), & r_1>p_{\mathrm{m}}\\ 0, & r_1<p_{\mathrm{m}}\end{cases},\tag{12}$$

$$\mathrm{pop}(j,k)=\begin{cases}\mathrm{pop}(j,k), & K=0\\ r_2(p_{\mathrm{popsize}}-1)+1, & K=1\\ r_2(p_{\mathrm{pop,max}}-p_{\mathrm{pop,min}})+p_{\mathrm{pop,min}}, & K=2\end{cases},\tag{13}$$

式中：$\mathrm{ceil}(2r_1)$ 表示将 $2r_1$ 四舍五入为大于或等于最接近 $2r_1$ 的整数；p_{popsize} 为种群最大数量；$p_{\mathrm{pop,max}}$、$p_{\mathrm{pop,min}}$ 分别为粒子群优化参数变化的最大值和最小值；K 为一个中间参数。

以 NH_3 气体回归预测模型 NH_3-SVM model 为例，具体优化步骤如图 6 所示。

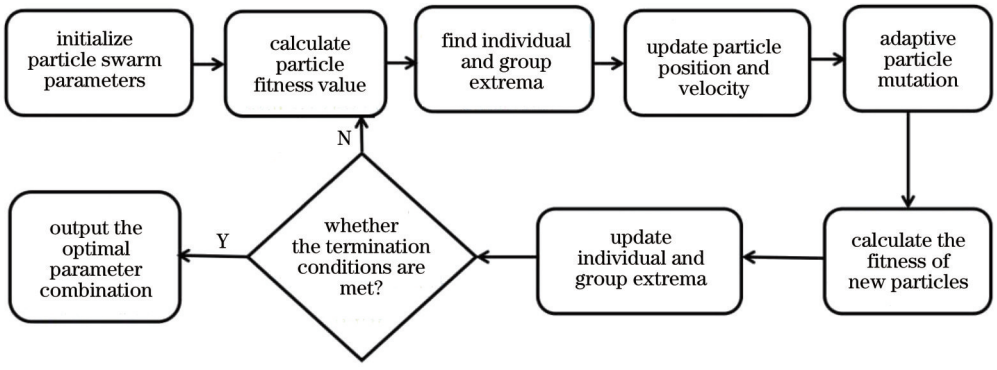

图 6　自适应变异粒子群优化算法流程图

Fig. 6　Flowchart of adaptive mutation particle swarm optimization algorithm

1）随机初始化模型参数。通过随机初始化模型参数得到 NH_3-SVM model 参数组合 (C,g) 在解空间的位置和个体粒子的初始速度和位置。每个粒子群只能优化一个模型参数，粒子群维数 $d=2$，粒子群 P_1 优化惩罚因子 $C\in(0.1,100)$，粒子群 P_2 优化核函数参数 $g\in(0.01,1000)$，惯性权重因子 $\omega=0.6$，学习因子 $c_1=1.5$、$c_2=1.7$，种群数量为 20 个，粒子群优化迭代次数为 200。

2）计算每个粒子的适应度。选择均方根误差作为目标函数 $F_{\mathrm{mse}}(C,g)$：

$$F_{\mathrm{mse}}(C,g)=\min\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[y_i-f(\boldsymbol{x}_i,C,g)\right]^2},\tag{14}$$

式中：$f(\boldsymbol{x}_i,C,g)$ 为适应度值。适应度函数为

$$F_{\mathrm{fit}}=F_{\mathrm{mse}}(C,g)。\tag{15}$$

设置每个粒子的初始位置为当前的个体极值p_{best}。根据式(15)得到每个粒子的个体适应度值$f_{present}$，若$f_{present} > p_{best}$，则更新个体最优值$p_{best} = f_{present}$，选取最大p_{best}更新为群体最优解g_{best}。

3) 粒子群更新。根据式(10)、(11)更新粒子位置和速度，并计算新的适应度值$f_{present}$。

4) 自适应变异。当粒子位置大于变异阈值p_m时，根据式(12)、(15)执行变异操作，更新适应度值$f_{present}$后，类似于步骤3)的操作，比较当前粒子的个体最优极值p_{best}和当前适应度值$f_{present}$的关系，得到群体最优解g_{best}。

5) 停止迭代判断。当满足上述比较条件或者迭代次数大于最大迭代次数200时，停止迭代并输出最优SVM模型参数组合(C, g)，否则继续迭代返回步骤2)。

4.4 结果分析

根据上述AMPSO算法流程，结合K折交叉验证算法[21]，本文K取值为6，将原始数据分成6个子集，针对每个子集数据分别继续一次测试，将剩下的5组子集数据作为训练集，通过验证获得这6个回归模型精度的平均值，作为交叉验证算法下的回归精度，分别得到NH_3-SVM model和CO_2-SVM model训练集的AMPSO算法适应度曲线，如图7所示，其中：横轴为优化代数，最大代数为200；纵轴为模型适应度值，即训练集样本的均方误差。由图7(a)可以看出，NH_3-SVM model训练集粒子的适应度值基本在0.1附近很小的范围内波动，迭代10次左右就可以得到最佳适应度，最终得到交叉验证下的均方误差为0.023；由图7(b)可以看出，CO_2-SVM model训练集粒子的适应度值在0.2～0.4范围内，迭代25次左右就可以得到最佳适应度，最终交叉验证下的均方误差为0.010。从均方误差的角度判断，两种气体回归预测模型的预测效果较好。

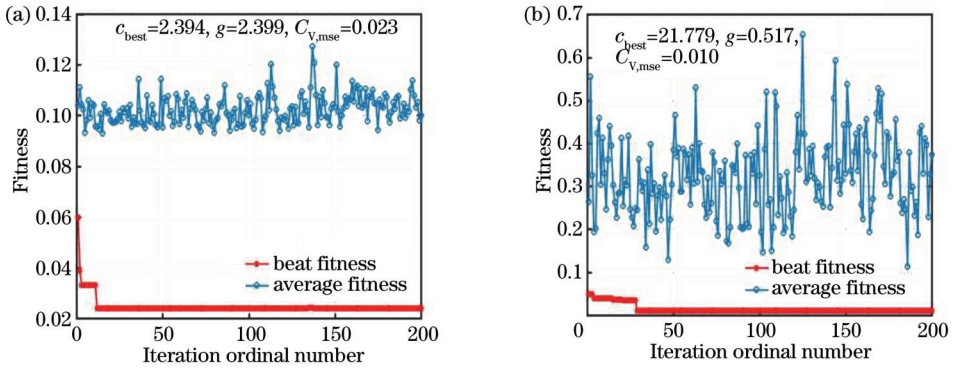

图7 模型训练集的自适应粒子群优化误差曲线。(a) NH_3-SVM model；(b) CO_2-SVM model

Fig. 7 Error curves of adaptive particle swarm optimization for model train set. (a) NH_3-SVM model; (b) CO_2-SVM model

为了对比AMPSO算法的优势，使用标准PSO算法和网格搜索法针对相同训练集的吸收光谱数据样本建立NH_3-SVM model和CO_2-SVM model，并进行结果对比分析。表2为采用三种不同寻优方法建立的SVM回归模型得到的最优参数组合(C, g)、均方误差(MSE)以及寻优时间。

表2 三种寻优算法的结果对比

Table 2 Comparison of results of three optimization algorithms

Model category	NH_3-SVM model			CO_2-SVM model		
	Parameter(C, g)	MSE	Optimization time /s	Parameter(C, g)	MSE	Optimization time /s
AMPSO	(2.394, 2.399)	0.023	29.367	(21.779, 0.517)	0.010	13.181
Standard PSO	(75.899, 0.010)	0.031	2.181	(0.100, 517.726)	0.052	1.199
Grid search	(0.732, 4.000)	0.024	39.661	(4.000, 0.758)	0.012	18.762

由表2可知：标准PSO算法虽然优化时间最短，但出现了过早成熟收敛的情况，导致均方误差较大，不能较好地对模型进行回归预测；网格搜索法虽然均方误差和AMPSO算法接近，都较小，但是该方法为非启发式算法，每次寻优需要遍历网格内所有点，导致寻优时间过长；AMPSO算法在较为合适的寻优时间下，可以得到最佳的均方误差，效率较高，这验证了AMPSO算法在构建混合气体体积分数回归模型方面的优势。

将AMPSO算法优化的NH_3-SVM model和CO_2-SVM model得到的最优参数组合(2.394, 2.399)和(21.779, 0.517)分别输入SVM，得到对应的体积分数回归模型。NH_3-SVM model和CO_2-SVM model训练集样本的预测结果如图8(a)、(b)所示，测试集样本的预测结果如图8(c)、(d)所示。

用决定系数R^2评估预测体积分数和设定体积分数的拟合情况。由图8可知：对训练集样本本身进行体积分数回归预测时，NH_3-SVM model和CO_2-SVM model的训练集体积分数设定值与体积分数预测值的

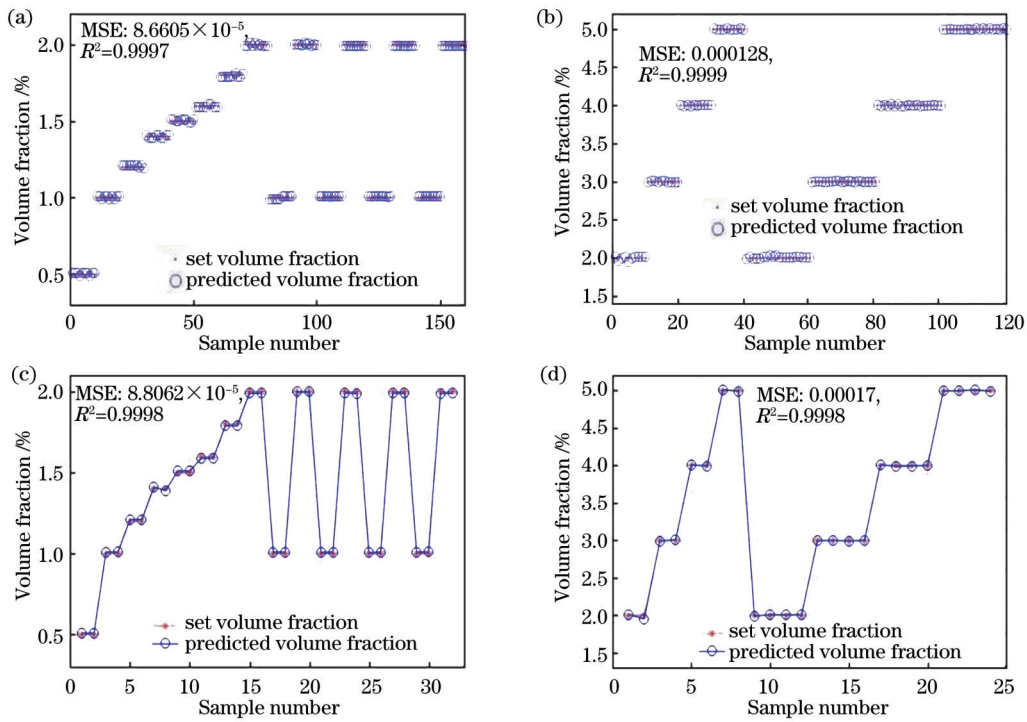

图 8 训练集和测试集结果对比。(a) NH_3-SVM model训练集设定体积分数和预测体积分数对比;(b) CO_2-SVM model训练集设定体积分数和预测体积分数对比;(c) NH_3-SVM model测试集设定体积分数和预测体积分数对比;(d) CO_2-SVM model测试集设定体积分数和预测体积分数对比

Fig. 8 Comparison of results of train set and test set. (a) Comparison of set volume fraction and predicted volume fraction for NH_3-SVM model train set; (b) comparison of set volume fraction and predicted volume fraction for CO_2-SVM model train set; (c) comparison of set volume fraction and predicted volume fraction for NH_3-SVM model test set; (d) comparison of set volume fraction and predicted volume fraction for CO_2-SVM model test set

均方误差分别为 0.000087 和 0.000128,决定系数 R^2 分别为 0.9997 和 0.9999;对测试集样本进行体积分数回归预测时,NH_3-SVM model 和 CO_2-SVM model 的测试集体积分数设定值与体积分数预测值的均方误差分别为 0.000088 和 0.000170,决定系数 R^2 均为 0.9998。结果表明:基于 AMPSO 算法建立的 NH_3 和 CO_2 气体体积分数回归预测模型的预测体积分数和实际体积分数可以很好地吻合,AMPSO 算法有较好的预测能力,拟合效果好,误差小。

由表 3 中基于 AMPSO 算法建立的 NH_3 和 CO_2 气体体积分数回归预测模型对于测试集气体体积分数的预测结果可以看出:NH_3 气体在设定体积分数为 1.6% 时,预测体积分数和设定体积分数之间的绝对误差最大,为 0.011%;在设定体积分数为 0.5% 时,预测体积分数和设定体积分数之间的相对误差最大,为 1.8%;编号为 16 的测试集的体积分数平均绝对误差为 0.008%,平均相对误差为 0.687%,最大相对误差在 2% 以内。CO_2 气体在设定体积分数为 2.0% 时,预测体积分数和设定体积分数之间的绝对误差和相对误差都最大,分别为 0.020% 和 1%,编号为 12 的测试集的体积分数的平均绝对误差为 0.005%,平均相对误差为 0.200%,最大相对误差在 1% 以内。综上,本文建立的 AMPSO 气体体积分数回归模型的预测精度高、准确性高,可应用于混合气体体积分数回归预测。

5 结 论

研究了一种基于自适应变异粒子群优化算法的混合气体体积分数定量分析方法,通过在标准粒子群优化算法的基础上引入自适应变异算子,使得寻优算法前期不易陷入局部最小值,避免了模型早熟,且算法后期收敛速度较快,提高了全局搜索能力。通过采用三种寻优方法分别对 NH_3 和 CO_2 气体训练集样本建立体积分数回归模型并进行对比分析,结果表明:基于 AMPSO 算法建立的 SVM 模型在较合理的寻优速度下,均方误差较小,模型回归效果最好;NH_3 和 CO_2 气体测试集设定体积分数和模型预测体积分数的绝对误差均在 0.02% 以内,NH_3 气体测试集设定体积分数和模型预测体积分数的相对误差在 2% 以内,CO_2 气体测试集设定体积分数和模型预测体积分数的相对误差在 1% 以内,预测精度较高。因此,基于自适应变异粒子群优化算法的 SVM 气体体积分数回归模型在对检测机动车排放的多组分混合气体的定量分析建模中具有一定的发展潜力和挖掘空间。

表 3 基于AMPSO的测试集结果

Table 3 Test set results based on AMPSO unit: %

Group	NH_3 volume fraction				CO_2 volume fraction			
	Set value	Predicted value	Absolute error value	Relative error value	Set value	Predicted value	Absolute error value	Relative error value
1	0.5	0.509	0.009	1.800	0	—	—	—
2	1.0	1.010	0.010	1.000	0	—	—	—
3	1.2	1.210	0.010	0.833	0	—	—	—
4	1.4	1.400	0	0	0	—	—	—
5	1.5	1.510	0.010	0.667	0	—	—	—
6	1.6	1.589	0.011	0.688	0	—	—	—
7	1.8	1.790	0.010	0.556	0	—	—	—
8	2.0	1.990	0.010	0.500	0	—	—	—
9	0	—	—	—	2.0	1.980	0.020	1.000
10	0	—	—	—	3.0	2.999	0.001	0.033
11	0	—	—	—	4.0	4.001	0.001	0.025
12	0	—	—	—	5.0	4.990	0.010	0.200
13	1.0	1.010	0.010	1.000	2.0	2.007	0.007	0.350
14	2.0	1.999	0.001	0.050	2.0	1.996	0.004	0.200
15	1.0	1.008	0.008	0.800	3.0	2.996	0.004	0.133
16	2.0	1.991	0.009	0.450	3.0	2.995	0.005	0.167
17	1.0	1.008	0.008	0.800	4.0	3.998	0.002	0.050
18	2.0	1.991	0.009	0.450	4.0	3.995	0.005	0.125
19	1.0	1.009	0.009	0.900	5.0	4.995	0.005	0.100
20	2.0	1.990	0.010	0.500	5.0	4.999	0.001	0.020

参 考 文 献

[1] Fraser M P, Cass G R. Detection of excess ammonia emissions from in-use vehicles and the implications for fine particle control [J]. Environmental Science & Technology, 1998, 32(8): 1053-1057.

[2] Pal D, Sen R, Pal A. Design of all-fiber thulium laser in CW and QCW mode of operation for medical use[J]. Physica Status Solidi C, 2017, 14(1/2): 1600127.

[3] Hardy L A, Wilson C R, Irby P B, et al. Rapid thulium fiber laser lithotripsy at pulse rates up to 500 Hz using a stone basket [J]. IEEE Journal of Selected Topics in Quantum Electronics, 2014, 20(5): 138-141.

[4] Li Z, Heidt A M, Simakov N, et al. Diode-pumped wideband thulium-doped fiber amplifiers for optical communications in the 1800-2050 nm window[J]. Optics Express, 2013, 21(22): 26450-26455.

[5] Gibert F, Flamant P H, Cuesta J, et al. Vertical 2-μm heterodyne differential absorption lidar measurements of mean CO_2 mixing ratio in the troposphere[J]. Journal of Atmospheric and Oceanic Technology, 2008, 25(9): 1477-1497.

[6] Bremer K, Pal A, Yao S, et al. Sensitive detection of CO_2 implementing tunable thulium-doped all-fiber laser[J]. Applied Optics, 2013, 52(17): 3957-3963.

[7] Al-Alawi S M, Abdul-Wahab S A, Bakheit C S. Combining principal component regression and artificial neural networks for more accurate predictions of ground-level ozone[J]. Environmental Modelling & Software, 2008, 23(4): 396-403.

[8] Kolehmainen M, Martikainen H, Hiltunen T, et al. Forecasting air quality parameters using hybrid neural network modelling[J]. Environmental Monitoring and Assessment, 2000, 65: 277-286.

[9] Kukkonen J, Partanen L, Karppinen A, et al. Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki[J]. Atmospheric Environment, 2003, 37(32): 4539-4550.

[10] 段小丽, 王明泉. 改进型PSO-SVM算法对井下多组分气体定量分析的研究[J]. 光谱学与光谱分析, 2019, 39(9): 2883-2888.
Duan X L, Wang M Q. Quantitative analysis of multi-component gases in underground by improved PSO-SVM algorithm[J]. Spectroscopy and Spectral Analysis, 2019, 39(9): 2883-2888.

[11] Zhang J W, Tittel F K, Gong L W, et al. Support vector machine modeling using particle swarm optimization approach for the retrieval of atmospheric ammonia concentrations[J]. Environmental Modeling & Assessment, 2016, 21(4): 531-546.

[12] 曲健, 陈红岩, 刘文贞, 等. 基于自适应变异粒子群优化的SVM在混合气体分析中的应用[J]. 传感技术学报, 2015, 28(8): 1262-1268.
Qu J, Chen H Y, Liu W Z, et al. Application of support vector machine based on adaptive mutation particle swarm optimization in analysis of gas mixture[J]. Chinese Journal of Sensors and Actuators, 2015, 28(8): 1262-1268.

[13] Zhang Y, Luo X G, He K, et al. Colorimetric artificial nose and pattern recognition methods for the concentration analysis of NH_3 [J]. Water, Air, & Soil Pollution, 2012, 223(6): 2969-2977.

[14] Soroush E, Shahsavari S, Mesbah M, et al. A robust predictive tool for estimating CO_2 solubility in potassium based amino acid salt solutions[J]. Chinese Journal of Chemical Engineering, 2018, 26(4): 740-746.

[15] Sedighi F, Vafakhah M, Javadi M R. Rainfall-runoff modeling using support vector machine in snow-affected watershed[J]. Arabian Journal for Science and Engineering, 2016, 41(10):

4065-4076.
[16] Alpert N L, Keiser W E, Szymanski H A, et al. IR-theory and practice of infrared spectroscopy[J]. Physics Today, 1974, 27(5): 47-49.
[17] Vapnik V N. An overview of statistical learning theory[J]. IEEE Transactions on Neural Networks, 1999, 10(5): 988-999.
[18] 盛文娟, 胡正彬, 杨宁, 等. 基于优化最小二乘支持向量机的温度稳定光纤布拉格光栅传感解调[J]. 激光与光电子学进展, 2022, 59(3): 0305002.
Sheng W J, Hu Z B, Yang N, et al. Demodulation of temperature stabilized fiber Bragg grating sensor based on optimized least square support vector machine[J]. Laser & Optoelectronics Progress, 2022, 59(3): 0305002.
[19] 张立欣, 张楠楠, 张晓. 基于机器学习算法对苹果产地的判别分析[J]. 激光与光电子学进展, 2022, 59(4): 0430001.
Zhang L X, Zhang N N, Zhang X. Discriminant analysis of apple origin based on machine learning algorithm[J]. Laser & Optoelectronics Progress, 2022, 59(4): 0430001.
[20] 朱霄珣, 韩中合. 基于PSO参数优化的LS-SVM风速预测方法研究[J]. 中国电机工程学报, 2016, 36(23): 6337-6342, 6598.
Zhu X X, Han Z H. Research on LS-SVM wind speed prediction method based on PSO[J]. Proceedings of the CSEE, 2016, 36(23): 6337-6342, 6598.
[21] 王其, 曾万聃, 夏志平, 等. 基于随机森林算法的食源性致病菌拉曼光谱识别[J]. 中国激光, 2021, 48(3): 0311002.
Wang Q, Zeng W D, Xia Z P, et al. Recognition of food-borne pathogenic bacteria by Raman spectroscopy based on random forest algorithm[J]. Chinese Journal of Lasers, 2021, 48(3): 0311002.

Application of Support Vector Machine in Quantitative Analysis of Mixed Gas

Shan Jifang[1,2], Liu Kun[1,2*], Jiang Junfeng[1,2], Liu Tiegen[1,2], Yin Hui[1,2]

[1]*School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin* 300072, *China*;
[2]*Key Laboratory of Opto-Electronics Information Technology, Ministry of Education, Tianjin University, Tianjin* 300072, *China*

Abstract

Objective Vehicle exhaust contains gases such as NH_3 and CO_2 and is becoming an essential source of air pollution and greenhouse effect. The intracavity absorption gas sensing technology based on fiber ring laser has many advantages, which are very suitable for real-time detection of toxic and harmful gases in environmental protection. However, when the gas sensing system based on a thulium-doped fiber laser is applied for quantitative analysis of mixed gas, the gas detection accuracy is often affected by cross interference caused by overlapping spectral absorption lines between component gases, and a nonlinear shift led to by changes in temperature and pressure at the experimental sites. As a small sample machine learning method, support vector machine (SVM) based on statistical theory has high accuracy and good generalization ability. It can be combined with infrared spectrum analysis to build a mixed gas volume fraction regression prediction model and correct nonlinear interference, thus greatly improving the accuracy and reliability of the gas quantitative analysis.

Methods In this paper, an active intracavity gas sensing system based on a thulium-doped fiber laser is built to collect the absorption spectrum data of NH_3 and CO_2 gases. The system is mainly divided into an adjustable light source (part A), a sensing part (part B), a data acquisition and processing part (part C), and a gas distribution part (part D). Before collecting the gas spectrum, sufficient nitrogen is introduced into the gas chamber to eliminate the interference of water vapor and CO_2 in the gas distribution instrument. The experimental environment is 0.1 MPa under normal pressure, and the sampling rate of the acquisition card is 20 kHz, with 20 groups of data being collected and 12 samples for each group of data. Before building the model, spectral data should be preprocessed to reduce the impact of background noise and improve the signal-to-noise ratio. However, it is inappropriate to do too much preprocessing to avoid losing some important spectral information. We also preprocess the spectral data through the methods of denoising, baseline correction, and smoothing. With an aim to improve the modeling speed, principal component analysis (PCA) is employed to project the multi-dimensional linear transformation of the original gas absorption spectrum data into a high-dimensional space to obtain the principal components corresponding to the maximum variance. The principal components at this time are leveraged to replace the eigenvalues in the original data, reduce the data dimension, and prevent the correlation between variables from affecting the extraction of these components and the prediction accuracy of the regression model. The standard particle swarm optimization (PSO) algorithm has fast convergence and short optimization time, whereas it features premature convergence of the model, low accuracy of optimal solution search, and low efficiency of later iteration. Therefore, we propose an improved algorithm, which is adaptive mutation particle swarm optimization (AMPSO). By

introducing an adaptive mutation operator, the updated particle positions are randomly mutated so that particles can enter other regions of the solution space to continue searching, thereby improving the ability of particle swarm optimization to jump out of the local optimal solution and avoid premature convergence of the algorithm model. The optimal combination of parameters obtained from the NH_3-SVM model and the CO_2-SVM model optimized by the AMPSO algorithm is input into the support vector machine to obtain the corresponding volume fraction regression model. The prediction results of training set samples and test set samples of the NH_3-SVM model and the CO_2-SVM model can be obtained (Fig. 8). The determination coefficient R^2 is adopted to evaluate the fit between the predicted volume fraction and set volume fraction.

Results and Discussions Although the optimization time of the standard PSO algorithm is the shortest, due to premature convergence, the mean square error is large, and the regression prediction of the model is not good. The mean square error of the grid search method is close to that of the AMPSO algorithm and both errors are small. However, since the grid search method is a non-heuristic algorithm, each optimization needs to traverse all points in the grid, resulting in long optimization time. Compared with the two algorithms, the AMPSO algorithm can obtain the best mean square error at a more appropriate optimization time, with higher efficiency. When regression predictions on the volume fraction of the training set samples are conducted, the mean square errors of the volume fraction set point and the volume fraction prediction value of the NH_3-SVM model and CO_2-SVM model are 0.000087 and 0.000128 respectively, and the determination coefficients R^2 are 0.9997 and 0.9999 respectively. When volume fraction regression prediction for the test set samples is carried out, the mean square errors of the volume fraction set point and the volume fraction prediction value of the NH_3-SVM model and CO_2-SVM model test set are 0.000088 and 0.000170 respectively, and R^2 is 0.9998.

Conclusions An active intracavity gas sensing system based on a thulium-doped fiber laser is built to collect the absorption spectrum data of NH_3 and CO_2 gases. The predicted volume fraction of the regression prediction model of NH_3 and CO_2 gas volume fraction is in good agreement with the actual volume fraction, with sound prediction ability and effect, and small error. The built AMPSO gas volume fraction regression model has high prediction accuracy and strong accuracy and can be applied for mixed gas volume fraction regression prediction.

Key words optical communications; thulium-doped fiber laser; adaptive mutation particle swarm optimization; mixed gas; support vector machine